Words in DNA Sequences : Some case studies based on their frequency statistics

Srabashi Basu
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta 700 035, India, srabashi@isical.ac.in.

Debi Prasad Burma
Professor Emeritus, Benaras Hindu University, Benaras, India.

Probal Chaudhuri
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta 700 035, India, probal@isical.ac.in.

1. Introduction

The invention of techniques for reading DNA sequences had a profound impact on the entire field of molecular biology. Over the last few years there has been a virtual explosion in the growth of data bases consisting of *large* DNA sequences. Integrated data bases are now globally accessible through the Internet. While in the early 80's the total volume of the world-wide DNA data bases was less than 10 million bases, by mid 90's this total volume has become nearly 200 million bases and continue to grow at an ever increasing pace. This has created the need for summarization of large volumes of sequence data so that effective statistical analysis can be carried out leading to fruitful scientific results. Various known sequence alignment algorithms and techniques for estimating the homologies and mis-matches among DNA sequences that are used for comparing sequences of relatively smaller sizes are not feasible to use when it comes to dealing with sequences with sizes varying between a few thousand base pairs to a few hundred thousand base pairs.

Our primary objective in this paper is to make a critical evaluation of the potentials of DNA word frequencies as a useful statistical summary of sequence data. It is expected that a careful analysis of word frequencies may also reveal valuable insights and interesting facts concerning the evolutionary process of different parts of a genome.

2. Structural Signature of a Genome Reflected in Word Frequencies

Usually prokaryotic organisms like bacteria and archaea have a single genome in their cells while a eukaryotic cell has several chromosomes. As the genome grew in size in course of evolution leading to complex cells, one plausible hypothesis is that the chromosomes were formed by fission of large genomes. In such a case, one would expect certain basic structural similarities in different chromosomes in the cell of an eukaryotic species. A related question is to what extent different parts of the same genome have similar structure.

2.1 An Analysis of the Complete Genome of Baker's Yeast

Yeast genome consists of sixteen chromosomes, and the smallest one among them is a sequence of 2,30,209 nucleotides while the largest one is a sequence of 15,83,176 nucleotides. Clearly, it is not possible to compare and judge the structural similarities of such large sequences with so much variations in their sizes using any of the standard sequence alignment and homology techniques. Besides, the biological functions of many parts of the genome are not well understood and certain parts of the genome might be biologically non-functional. This makes total comparison of these sixteen chromosomes through functional homology virtually impossible. However, it is possible to compare the frequencies of various DNA words in these chromosomes, and this can lead to an idea of the extent of their structural similarities.

The first step towards that direction is a visual inspection of the frequencies for 2-, 3-, 4- and 5-words in all sixteen chromosomes. There is a high degree of similarity in the word frequencies for all those chromosomes. A note-worthy fact is that each of the 4^6 6-words occurs with positive frequency in each of the chromosomes, and in each chromosome, some of the 4^8 8-words do not occur. In each of the seven larger chromosomes, all of the 4^7 7-words occur with positive frequency, while in each of the nine smaller chromosomes, some of the 7-words turn out to be missing.

In order to make a critical comparison of all sixteen chromosomes, it will be appropriate to consider relative ranks of different k-words in the chromosomes in terms of their frequencies. Strong structural similarity between two sequences, which is reflected in relative abundance and scarcity of different DNA k-words in the two sequences, will make the value of correlation coefficient very close to $+1$.

Next we make an attempt to investigate to what extent 2-word frequencies are reflected in k-word frequencies for $k \geq 3$. We generate sixteen random sequences and a 1-step homogeneous Markov process whose 1-step transition probabilities are estimated from the 2-word frequencies of a chromosome. Then we can compare the k-word frequencies for sixteen random sequences with those for sixteen actual chromosomes for $k \geq 3$. Let α^* denote the random Markov sequence generated using the 2-word frequencies of the chromosome α. For each of the sixteen chromosomes, we will make the sequence α^* to have the same length as α, and both will begin with the same letter. It is clearly noticeable that the values of cross-correlations of ranks for k-words of actual chromosomes and random Markov sequences are consistently smaller than the corresponding values of pairwise rank correlations among actual chromosomes. This is an evidence for the fact that k-words for $k \geq 3$ have their intrinsic frequency patterns in each of these chromosomes that are not derivable from their 2-word frequencies. This analysis on the one hand, proves the existence of a very strong structural similarity of the sixteen chromosomes, and on the other hand, it reveals that the frequencies of higher order words in this case have their own patterns that are not completely derived from lower order word frequencies.

2.2 Analysis of Some Bacteriophage Genomes

We have chosen four bacteriophage genomes : $\Phi X174$ (5,386 nucleotides), $G4$ (5,577 nucleotides), $F1$ (6,407 nucleotides) and $PF3$ (5,833 nucleotides) and have considered five fragments of each of these bacteriophage genomes. The first two fragments are formed by taking two equal and disjoint halves of the entire genome, and the other three fragments are formed by taking three equal and disjoint parts of the full genome. Average linkage cluster analysis based on simple l_1-distances of 3- and 4-word frequencies of twenty fragments are performed. Here the l_1-distance between a pair of frequency vectors is given by $\sum_w |f_w - g_w|$, where f_w's and g_w's are the frequencies corresponding to 3-words (or 4-words) in two fragments of the same (or different) phage genome(s). In both the dendrograms, fragments of the same bacteriophage genome exhibit an overall tendency of forming their own separate clusters, and this is indicative of the presence of some form of structural signature in different fragments of the same phage that is amply captured by the word frequencies.

An important issue that arises at this point is how to identify those k-words whose frequencies bear stronger impression of the structural signature of any specific virus genome compared to other k-words. Given n different phage genomes and m fragments of each genome, let $f_w^{a,\alpha}$ denote the frequency of the k-word w in the fragment α of the phage a. Define now the ratio

$$F(w) = \frac{m^{-1}(n-1)^{-1}\sum\limits_{a \neq b}\sum\limits_{\alpha,\beta} |f_w^{a,\alpha} - f_w^{b,\beta}|}{(m-1)^{-1}\sum\limits_{a}\sum\limits_{\alpha \neq \beta} |f_w^{a,\alpha} - f_w^{a,\beta}|},$$

which may be interpreted as the l_1-distance version of the traditional F-ratio of between and within group variations used in the analysis of variance. A very large value of $F(w)$ corresponds to significantly larger value of the dispersion among the frequencies $f_w^{a,\alpha}$ across different phages (*i.e.* a's) compared to the dispersion among these frequencies across different fragments (*i.e.* α's) of the same phage. One can then rank the k-words according to the F-values, and we will call it the F-rank of the k-word w. It seems meaningful to choose k-words with high F-ranks in order to capture structural signatures of different virus genomes.

In the case of twenty fragments of the four phage genomes that we have formed, we selected fifteen 3-words and fifteen 4-words that have highest F-ranks among all 3-words and 4-words respectively. Then these word frequencies were utilized to carry out Chernoff's face analysis. In both cases, the similarity among the faces in each column corresponding to different fragments of the same virus genome is quite noticeable, and so are the dissimilarities of the faces in different columns that correspond to different virus genomes.

3. Word Frequencies and Phylogenetic Relationships

Phylogenetic relationships among different organisms are of fundamental importance in biology, and one of the prime objectives of DNA sequence analysis is the construction of phylogenetic trees for understanding evolutionary history of organisms. We have analyzed 5S RNA data for ten species using the l_1-distance of frequency vectors. The size of the 5S RNA sequences in the data set vary between 118 and 122 nucleotides. Clustering was carried out in a hierarchical aglomerative manner using single and average linkage on trinucleotide frequencies. The dendrograms show biologically meaningful clustering, where fungi, bacteria, plants, animals and chloroplasts form distinct groups, and the eukaryotes nicely separate out from prokaryotes and chloroplasts.

Our next data set consists of 16S and 18S ribosomal RNA sequences for twenty organisms. These sequences have their sizes varying between 1471 and 1869 nucleotides. We observe from the results of cluster analysis based on 4-word frequencies that the bacteria, the archaea, the fungi and the mammals form distinct clusters, and the eukaryotes clearly separate out from the prokaryotes.

It is now clear that different biological species exhibit different distributions of DNA words in their ribosomal RNA sequences, and this can be amply utilized to cluster different organisms into homogenous biological groups by a straight-forward comparison of their DNA word frequencies. It appears that one can use distances based on DNA word frequencies as an alternative and effective statistical tool for quick and convenient phylogenetic and taxonomic analysis.

4. Concluding Remarks and Discussion

If k-word frequencies are to be used to summarize a given DNA sequence, a question that naturally arises is what value(s) of k one should use. There does not seem to be a general mathematical solution to this problem, and one has to settle it depending on the specific situation. For some case studies 3- and 4-word frequencies yield results that are in good agreement with known biological facts. For very large sequences like the chromosomes in yeast, one would probably go further beyond 4-words. Note that if the total length of a sequence is N, for any $k > \log_4(N - k + 1)$ (i.e. $4^k > N - k + 1$), some of the k-words are bound to be missing in that sequence. Since $N >> k$, we will have $\log_4(N - k + 1) \approx \log_4 N$, and it appears that the frequencies of words having size larger than $\log_4 N$ may not be statistically very informative due to the "sparsity" of such words in the sequence. For the sixteen chromosomes in yeast, the values of $\log_4 N$ vary between 8.91 and 10.30. We have already pointed out that each of the chromosomes has some of the 8-words missing, and each of the nine smaller chromosomes has even some of the 7-words missing in their sequences. We feel that $\log_4 N$ can be used as a preliminary empirical guideline for choosing appropriate value(s) of k for carrying out meaningful statistical analysis based on k-word frequencies. Upon exploration of some relations among the frequencies of k-words for different values of k, it was found that the frequency of any $(k-1)$-word is almost equal to a weighted sum of the k-word frequencies, where the weights are equal to either 0 or 1. Generalizing further, we may claim that for $N >> k > k^* \geq 1$, different k^*-word frequency can be well approximated by different orthogonal weighted sums of k-word frequencies where the weights are $0 - 1$ valued.

Word frequencies are very naive yet quite natural and useful statistical summaries of DNA sequences. In view of the massive size of DNA sequence data that has been made available these days by automated biotechnology, there is a real need for statistical summarization to gain information from this kind of data. It is fairly transparent from our data analysis that one can conveniently use word frequencies to capture structural patterns present in the DNA sequences of the same organism or organisms forming homogenous biological groups. We have also observed that frequency distributions for DNA words tend to be different in the sequences obtained from different biological groups. All these can make such relatively simple statistical summary measures very useful tools for analysis of large DNA sequences, which is currently an important and interesting problem that lies in the interface between statistics and molecular biology.